

# Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse

Jonathan C. Prather, M.S.<sup>1</sup>, David F. Lobach, M.D., Ph.D., M.S.<sup>1</sup>, Linda K. Goodwin, R.N., Ph.D.<sup>2</sup>, Joseph W. Hales, Ph.D.<sup>1</sup>, Marvin L. Hage, M.D.<sup>3</sup>, and W. Edward Hammond, Ph.D.<sup>1</sup>

Duke University Medical Center, Durham, North Carolina

<sup>1</sup>Division of Medical Informatics

<sup>2</sup>School of Nursing

<sup>3</sup>Department of Obstetrics and Gynecology

*Clinical databases have accumulated large quantities of information about patients and their medical conditions. Relationships and patterns within this data could provide new medical knowledge. Unfortunately, few methodologies have been developed and applied to discover this hidden knowledge. In this study, the techniques of data mining (also known as Knowledge Discovery in Databases) were used to search for relationships in a large clinical database. Specifically, data accumulated on 3,902 obstetrical patients were evaluated for factors potentially contributing to preterm birth using exploratory factor analysis. Three factors were identified by the investigators for further exploration. This paper describes the processes involved in mining a clinical database including data warehousing, data query and cleaning, and data analysis.*

## INTRODUCTION

Vast quantities of data are generated through the health care process. While technological advancements in the form of computer-based patient record software and personal computer hardware are making the collection of and access to health care data more manageable, few tools exist to evaluate and analyze this clinical data after it has been captured and stored. Evaluation of stored clinical data may lead to the discovery of trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management. Techniques are needed to search large quantities of clinical data for these patterns and relationships. Past efforts in this area have been limited primarily to epidemiological studies on administrative and claims databases. These data sources lack the richness of information that is available in databases comprised of actual clinical data. In this study we propose the introduction of a recently developed methodology known as data mining to clinical databases.

Data mining, also referred to as Knowledge Discovery in Databases or KDD, is the search for relationships and global patterns that exist in large databases but are 'hidden' among the vast amounts of data [1]. The typical data mining process involves transferring data originally collected in production systems into a data warehouse, cleaning or scrubbing the data to remove errors and check for consistency of formats, and then searching the data using statistical queries, neural networks, or other machine learning methods [2]. Most previous applications of KDD have focused on discovering novel data patterns to solve business related problems such as designing investment strategies or developing marketing campaigns.

Data warehousing and mining techniques have rarely been applied to health care. Recently, researchers at the Southern California Spinal Disorders Hospital in Los Angeles used data mining to discover subtle factors affecting the success and failure of back surgery which led to improvements in care [3]. In a second health care application, GTE Laboratories built a large data mining system that evaluated health-care utilization to identify intervention strategies that were likely to cut costs [3]. This system, however, is focused on cost analysis and not on identifying new associations or relationships within clinical data.

We are currently in the process of initiating a data mining project at Duke University Medical Center using an extensive clinical database of obstetrical patients to identify factors that contribute to perinatal outcomes. The purpose of this paper is to illustrate how medical production systems such as the Duke Perinatal Database can be warehoused and mined for knowledge discovery. The eventual goal of this knowledge discovery effort is to identify factors that will improve the quality and cost effectiveness of perinatal care.

**Table 1.** Characteristics of the production system database and the clinical data warehouse.

Characteristic	Production System Database	Clinical Data Warehouse
Physical Size	<i>265 megabytes</i>	<i>845 megabytes</i>
Database Structure	<i>Node-based class-oriented</i>	<i>Relational: Microsoft SQL Server Version 4.2</i>
Clients	<i>VT-420, Telnet</i>	<i>Any ODBC or db-lib compliant interface</i>
Hardware Platform	<i>VAX 6000-230 Mini Computer</i>	<i>PC with a 60 MHz Pentium CPU, 16 mb RAM</i>
Performance optimization	<i>Single record retrieval</i>	<i>Cross-population queries</i>
Query Mechanism	<i>TMR program code Version 9.0</i>	<i>Structured Query Language (SQL)</i>
Operating System	<i>VMS Version 5.5</i>	<i>Windows NT™ Server Version 3.5</i>

In this paper we describe the medical data mining process which entails transfer of the database from a comprehensive computer-based patient record system (CPRS) into a data warehouse server, creation of a dataset for analysis by extracting and cleaning selected variables, and mining of the data using exploratory factor analysis. We report the preliminary results of factor analysis on two years of perinatal data and compare these results with other studies in the field. Finally, we discuss several issues that should be considered in warehousing clinical data for mining.

## METHODS

### Production System Database

The production system database identified for mining was the computer-based patient record system known as The Medical Record, or TMR. TMR is a comprehensive longitudinal CPRS developed at Duke University over the last 25 years. The data collected in TMR include demographics, study results, problems, therapies, allergies, subjective and physical findings, and encounter summaries. TMR's data structure uses a proprietary class-oriented approach which stores all of the patient's information in a single record.

The specific TMR database selected for this project was the perinatal database used by the Department of Obstetrics and Gynecology at Duke University Medical Center. This database continues to serve as the repository for a regional perinatal computerized patient record that is used in inpatient and outpatient settings [4]. The on-line Duke perinatal database contains comprehensive data on over 45,000 unique patients collected over nearly 10 years. Additional patient data from the previous decade is also available on tape archive. This computerized repository contains more than 4,000 clinical variables collected on over 20,000 pregnancies and births from a five county area, making it one of the

largest and most comprehensive obstetrical datasets available for analysis in the United States.

### Creating the Data Warehouse

The data warehouse was created on a centralized server dedicated to fielding data mining queries. Using a method previously described [5], the clinical data was mapped from the proprietary TMR data structure into relational tables in the personal computer environment. Microsoft SQL Server V 4.2 was chosen as the database engine [6] and was installed on a PC server with a 60 MHz Pentium CPU, 1700 megabytes of hard disk, 16 megabytes of RAM, and using the Windows NT™ Server 3.5 operating system and file system. A comparison of the production system database and the clinical data warehouse is shown in Table 1.

### Extracting and Cleaning the Dataset for Analysis

For the purposes of this study, a sample two-year dataset (1993-1994) from the data warehouse was created to be mined for knowledge discovery. Multiple SQL queries were run on the data warehouse to create the dataset. As each variable was added to the dataset, it was cleansed of erroneous values, data inconsistencies, and formatting discrepancies. This cleaning process was accomplished using Paradox Application Language scripts to selectively identify problems and correct the errors.

The crucial role of these scripts was to scan the dataset and convert alphanumeric fields into numerical variables in order to permit statistical analysis. After checking to see if data values were collected during or pertaining to the preterm course of the infant, the script ensured that multiple values for the same variable were not present. If such values existed, the value that was recorded closest to delivery or conception, depending on perceived data quality for the particular variable, was loaded into the final dataset. A final script identified missing

**Table 2. Unusable data values encountered while extracting and cleaning the dataset variables for analysis.**

Reason Unusable	Example	Count	% of Total Values
Missing values when required	<i>Ward clerk did not enter or data item was not collected</i>	2,213	2.95%
Incomplete dates	<i>Dates preventing calculations, e.g. ??/??/94</i>	249	.33%
Free-text in place of a coded data phrase	<i>Ward clerk enters free text for an item in place of a code from the data dictionary</i>	4,071	5.43%
Other errors	<i>Out of range values, format discrepancies, data inconsistencies</i>	16	.02%
<b>Totals:</b>		<b>6,549</b>	<b>8.74%</b>

values and prompted the user to either substitute them with an average value for the variable, or to delete the subject from the dataset.

Robust demographic variables such as age, race, education, and marital status were automatically selected for inclusion in the dataset, while other routinely collected data elements were randomly selected for inclusion in the study. These elements originated in the problem section and the subjective and physical findings section of the electronic patient records.

#### Mining the Dataset

For this preliminary study, we selected exploratory factor analysis for data mining because it had previously been used successfully to explore claims and financial databases in obstetrics [7].

Factor analysis is a statistical method used to identify which data elements can be combined to explain variations between patient groups. This mining technique is appropriate in research problems in which a large number of subjects are compared on a set of variables for which there is no designation of independence or dependence [8].

The statistical software used to conduct the factor analysis was SPSS for Windows version 5.0, SPSS Inc., Chicago, IL.

## RESULTS

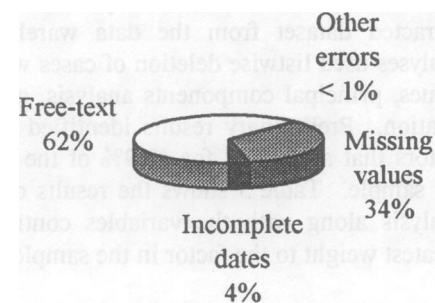
#### Creating the Clinical Data Warehouse

Following successful transfer of the entire production system database into the SQL Server data warehouse in the personal computer environment, the warehouse contained 45,922 patient records. These records included 215,626 encounters; 1,757,118 historical data elements; 3,898,887 individual lab

results; 217,453 problems and procedures; and 3,016,313 subjective and physical findings.

#### Extracting and Cleansing a Test Dataset

The average speed of the queries directed against the data warehouse was approximately 3 minutes, while the longest query for the study required 12 minutes to complete and occurred against a table of nearly 4 million records. The test dataset extracted from the data warehouse contained data regarding 3902 births occurring between January 1, 1993, and December 31, 1994.



**Figure 1. Causes of unusable data identified during extraction and cleansing of the dataset.**

The data cleaning programs used in creating the dataset revealed that 8.74% of the total values in the database were unusable for the purposes of the factor analysis. A breakdown of reasons why data could not be used, the counts of unusable value occurrences, and the percentages of the total values for each reason are shown in Table 2. Free text stored in place of a coded data phrase accounted for the largest amount of unusable data; missing values accounted for the second largest; and incomplete dates, the third largest. Other causes of unusable

**Table 3. Results of exploratory factor analysis of preterm delivery showing factors in the sample dataset.**

Warehoused Clinical Variable	Factor 1:	Factor 2:	Factor 3:
Variable 1	-.00750	.11454	.49309
Variable 2	-.45976	.08934	.10848
Variable 3	.10009	-.06756	.61430
Variable 4	-.81037	-.01755	-.04758
Variable 5	.73141	-.02381	.04549
Variable 6	.72223	.00688	.07664
Variable 7	.11447	-.79700	.07708
Variable 8	-.07415	-.07019	.67161
Variable 9	.02997	.82262	.07137

data included out of range values such as invalid heights, format discrepancies such as a date in a numeric field, and data inconsistencies such as two very different heights for one patient, combined in one group. The relative proportions of unusable data contributed by each of the above reasons is depicted in Figure 1. While unusable for analysis, the option to store free-text and incomplete dates represent legal values and are not true data errors. Thus, only 35% of the unusable values were actually caused by erroneous data.

#### **Mining the Data: Exploratory Factor Analysis**

Factor analysis was successfully conducted on the extracted dataset from the data warehouse. All analyses used listwise deletion of cases with missing values, principal components analysis, and varimax rotation. Preliminary results identified three latent factors that accounted for 48.9% of the variance in the sample. Table 3 shows the results of the factor analysis along with the variables contributing the greatest weight to the factor in the sample dataset.

We characterized three preliminary factors from the limited dataset, and plans are in place to mine a more extensive subset of data from the warehouse to improve the mining process and the identifications of factors of preterm birth risk.

### **DISCUSSION**

In this preliminary study we demonstrated that a large clinical database could be successfully warehoused and mined to identify clinical factors associated with preterm birth, even on a preliminary dataset. We also observed that the majority of unusable data in a sample dataset represented legal values that were excluded because they were not acceptable by our data mining approach of factor analysis.

The quality and comprehensiveness of the warehoused data is encouraging in comparison to previous studies on obstetrical databases. One such study found only 9 subjects out of 6,616 could be analyzed without missing fundamental values [7]. The investigators of this study noted that obstetrical databases they evaluated contained erroneous, poorly organized, inconsistently recorded and frequently dichotomous data elements. They also observed important data that might be associated with preterm birth were often not collected. These data included elements about stress, sexual activity, substance abuse, nutritional status, and infections, among others [7].

The data warehouse at Duke, however, contains comprehensive, quality clinical data on a large volume of patients. Exploratory factor analysis on a limited dataset revealed several variables which could significantly help categorize patients (variable weights greater than .4 and lower than -.4 are typically considered significant in factor analysis).

Newly discovered relationships found in clinical databases such as these will potentially lead to better understanding between observations and outcomes in perinatal care and other fields of medicine. In spite of numerous risk scoring tools and preterm birth prevention programs initiated in the 1980's, prematurity remains the most common cause of low birthweight and associated morbidity and mortality. The inadequacy of current risk scoring tools and preterm birth prevention programs stems from a failure to fully identify factors that cause preterm birth. With a more robust model of preterm delivery there exists the possibility of more specific prospective trials of strategies for prevention. Computerized patient record systems (CPRS) now contain enough detailed clinical data to help us find

the relationships between multiple perinatal variables and the outcomes they influence.

Data warehousing and mining technology are applicable to health care, and the preliminary mining of a clinical data warehouse has produced promising results. The new paradigm proposed in this paper for determining complex associations which influence medical outcomes by combining data mining with the computerized patient record merits further study. By implementing CPRS data warehousing, new medical hypotheses can be generated for predicting and preventing preterm birth and other adverse health outcomes.

*This work supported in part by NLM Training Grant #LM07071-3 and AHCPR Dissertation Grant #R03 HS09331-01.*

## References

1. Holsheimer M and Siebes A. (1994) Data Mining: the search for Knowledge in Databases. *Technical report CS-R9406*, CWI, January.
2. Krivda CD. (1995) Data-Mining Dynamite. *Byte*, Oct 95:97-102.
3. Hedberg, SR. (1995) The Data Gold Rush. *Byte*, 1995;Oct :83-88.
4. Burkett, ME. (1989) The Tertiary Center and Health Departments in Cooperation: The Duke University Experience. *J Perinat Neonat Nursing*, 2:11-19.
5. Prather JC, Lobach DF, Hales JW, Hage ML, Fehrs SJ, Hammond WE. (1995) Converting a Legacy System Database into Relational Format to Enhance Query Efficiency. *Proceedings Annual Symposium Computer Applications Medical Care*, 19:372-376.
6. Greenfield L. (1996) The Data Warehousing Information Center. LGI Systems Incorporated. URL: <http://pwp.starnetinc.com/larryg/database.html>
7. Woolery, L, and Grzymala-Busse, J. (1994) Machine learning and preterm birth risk assessment. *Journal of the American Medical Informatics Association*. 1(6):439-446.
8. Dawson-Saunders B, Trapp RG. (1994) *Basic and Clinical Biostatistics, Second Edition*. Appleton and Lange, 227-228.